Health Care Predictive Analytics using Artificial Intelligence Techniques

Guanjin Wang

A thesis submitted for the Degree of

Doctor of Philosophy

Faculty of Engineering and Information Technology

University of Technology Sydney

July 2018

Certificate of Originality

This thesis is the result of a research candidature conducted jointly with another University as part of a collaborative Doctoral degree. This research is also supported by the Australian Government Research Training Program. I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgment has been made in the text.

Student: Guanjin Wang

Date: 25/07/2018

Abstract

In recent years, advances in Artificial Intelligence (AI) are opening the door for intelligent health care data prediction and decision making. Machine learning, as an increasingly popular approach to AI, has been widely used to learn directly from data, adapt independently, and produce predictive outcomes, which support doctors when encountering complex health care predictive analytics. However, traditional machine learning methods are not always perfectly working in the health field, intrinsically due to little consideration for characteristic problems within health care data. For example, the small sample size problem is common due to complex data collection procedures and privacy concerns. Missing data is also widely encountered since most data are collected as a second-product of patient-care activities instead of following systematic research protocols. The class imbalance is another inevitable problem in the medical data as the normal class always predominates over the disease class. To solve aforementioned issues in health care predictive analytics, this study stands on the principles of machine learning and transfer learning to develop five advanced prediction models.

The first model is an output-based transfer least squares support vector machines (LS-SVMs) model which can leverage knowledge from the existing prediction model or on-line tool to facilitate the learning process on the current domain of interest with insufficient data. This model

overcomes the small sample size problem and improves the health care data prediction by learning knowledge from the other domain.

The second model is a novel additive LS-SVMs model which can make predictions simultaneously considering the influences on the classification error caused by missing features in a dataset. This model can generate valuable explanations regarding the influence levels of missing features for health professionals to improve the future data collection process.

The third model is a transfer-based additive LS-SVMs model which can deal with missing data from a transfer learning perspective. It can leverage the model knowledge learned from the complete portion of the dataset to help the learning process on the whole dataset with missing data. The proposed model can provide supplementary information for health professionals to improve the data quality via data cleaning.

The forth model is a deep transfer additive LS-SVMs model called DTA-LS-SVMs and its imbalanced version called iDTA-LS-SVMs to enhance the prediction performance on the balanced and imblanced datasets. Inspired by the stacked architecture and transfer learning mechanism, the model stacks multiple additive LS-SVMs based modules layer-by-layer and embeds model transfer between adjacent modules to guarantee their consistency.

The fifth model is a deep cross-output transfer LS-SVMs model called DCOT-LS-SVMs and its imbalanced version called IDCOT-LS-SVMs to improve the prediction performance on the balanced and imbalanced datasets. The cross-output transfer is used to transfer the predictive outcome from the previous module to the adjacent higher layer to achieve a better learning. Moreover, modules' parameters can be randomly assigned

in the proposed model which significantly reduces the time for model selection.

The proposed models are verified using experiments on the public UCI datasets. Moreover, case studies are conducted to validate and integrate the proposed models with real world applications, including bladder cancer prognosis, prostate cancer diagnosis, and predictions of elderly quality of life (QOL). The results have demonstrated that the models in this study can enhance the prediction performance while taking the characteristic problems within health care data into account, thus exhibiting promising potential for use in different health applications in future.

Acknowledgements

PhD study has been a challenging and memorial journey in the past three years. I would like to extend my warmest gratitude to the people who inspired and helped me along this journey.

First, I would like to express my earnest thanks to my chief supervisor in the Hong Kong Polytechnic University, Professor Choi Kup Sze, and my chief supervisor in the University of Technology Sydney, Distinguished Professor Jie Lu for their continuous support, precious guidance and enormous patience throughout my study. Without their excellent supervision and valuable suggestions, this Joint-PhD study between two universities could not have been finished. Thank you for all your detailed comments and suggestions on my research, manuscripts and presentations. Your strict academic attitude and hard working style have deeply influenced me and will alway inspire and motivate me in my future work and life. I also would like to address my sincere thanks to Professor Guangquan Zhang for his academic suggestions and advices, Dr Chiang Chung Lim Vico for his health care knowledge sharing and advices and Dr Lam Kin Man for his medical knowledge sharing and data support.

I am honored to have met all the great researchers and staffs in the Centre for Smart Health in the Hong Kong Polytechnic University and Centre for Artificial Intelligence in the University of Technology Sydney. I appreciate all their valuable and critical comments during my presentations, and the discussions with them were enlightening.

I am grateful to the School of Nursing in the Hong Kong Polytechnic University and the School of Software in the Faculty of Engineering and Information Technology in the University of Technology Sydney. This study was supported by the YC Yu Scholarship for the Centre for Smart Health, UTS Doctoral Scholarship, the Research Grants Council of the Hong Kong SAR and the Australian Research Council (ARC) discovery project.

Lastly, I would like to express my most grateful appreciation to my family for their unconditional love, encouragement and support. This journey would not have been possible without their help. I am especially grateful to my parents. Thank you for being the first teacher in my life and guiding me to be a better person. I always know you believe in me more than I do in myself.

Contents

List of Figures

List of Tables